Music Genre Classification using Convolutional Recurrent Neural Networks

Noopur Srivastava, Shivam Ruhil, Gaurav Kaushal
ABV-Indian Institute of Information Technology & Management, Gwalior, India
Email: noopurs@iiitm.ac.in, shivamruhilh@gmail.com, kaushalg@iiitm.ac.in

2022 IEEE 6th Conference on Information and Communication Technology (CICT) | 978-1-6654-7312-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CICT56698.2022.9997961

***Abstract*—Music genre classification is the task of classifying audio clips into well defined music genres. It is a popular task in the field of deep learning. We use CRNNs (Convolutional Recurrent Neural Networks) for the task. CRNNs are able to take advantage of both the spatial features and the temporal features of the data. We use MFCCs as a representation of our audio data. We use two models of CRNN : CNN-GRU and CNN-LSTM. We achieve the accuracy of 85.7% using CNN-GRU and an accuracy of 87.5% using CNN-LSTM on the GTZAN dataset.**

***Index Terms*—music genre classification, convolutional neural networks, recurrent neural networks, convolutional recurrent neural networks.**

I. Introduction

Digital audio files are uploaded to the Internet in enormous quantities due to the quick development of multimedia technologies. In addition to the advantages these audio activities offer, their rapid expansion has effects on a number of fronts. As a result, properly handling these audios is a difficult undertaking in need of solid solutions. A lot of effort has been put into research in Music Information Retrieval (MIR) to deal with various audio tasks. With the growth of apps such as the music recommendation systems and music search there has been an increasing interest in MIR. A very core issue of MIR is classifying the music into various genres. Content-based genre recognition is extremely important to bootstrap MIR system because expert annotation is notoriously expensive and difficult for huge catalogues.

Music Genre Classification is a popular task in the field of Deep Learning. [18] introduced the GTZAN dataset which has since been used as a benchmark for the problem. [11] introduced the Mel frequency analysis and cepstrum analysis for processing audio and extracting useful features for audio. [16] tried to focus on feature extraction and used tried traditional machine learning models such as SVM, K-NN, GMM with those features. Recently, the focus has shifted from feature extraction to deep learning. Convolutional Neural Networks have been explored extensively in many research works. [9] used Convolutional Neural Networks for Music Genre Classification on GTZAN dataset. [8] showed that RNN-LSTMs also work good for audio data. Some research papers have also used RNN-LSTM models in combination with other models like SVM on the same task [15].

Even though academics have put forth a variety of methods from different angles, the majority of them depend on creating great hand-crafted features and suitable classifiers for music data. Many of the existing research work has been directly utilizing features such as MFCCs and mel-spectrograms for various models such as Convolutional Neural netowrks. [9] uses CNN with mel-spectrograms. Mel Frequency Cepstral Coefficients (MFCC) are a good representation of audio data and capture the short term power spectrum of a sound. Through convolutional and pooling operations, CNN extracts the spatial and hierarchical features from an image. Similar to images music also has hierarchical structures. Therefore it seems intuitive to use CNNs for music genre classifications as well. However, the models using CNNs are not able to

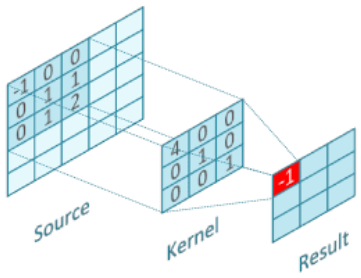

Fig. 1. Convolution operation in CNN [2]

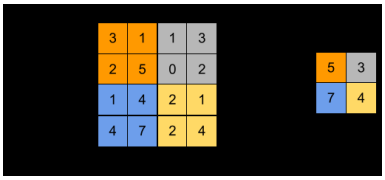

Fig. 2. Pooling operation in CNN.[4]

capture the long term temporal information of music data. Recurrent Neural Networks (RNN) are neural networks where connections between nodes form a sequence. RNNs are neural networks with loops and are great with sequential and time series data. Therefore, RNNs are a good way to capture the temporal relationships and therefore it is also intuitive to involve RNNs in our model for music genre classification. We have used Convolutional Recurrent Neural Networks (CRNN)

978-1-6654-7312-5/22/$31.00 ©2022 IEEE

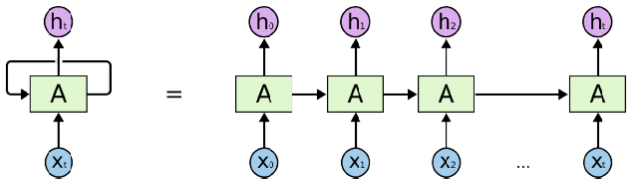

Fig. 3. An unrolled loop in RNN[7]

which are a combination of CNN and RNN. CRNNs have proven to be successful in similar tasks [10]. In this hybrid model, CNN helps extract spatial features and RNN helps derive the temporal relationships in data. The models with CNN [9] don't utilise the temporal features and only using LSTM won't take the advantage of the hierarchical features of the music clip. [10] used the idea of CRNN (CNN-GRU) for the problem of music tagging and achieved improved results. So the CRNN model should also give promising results on the task of music genre classification.

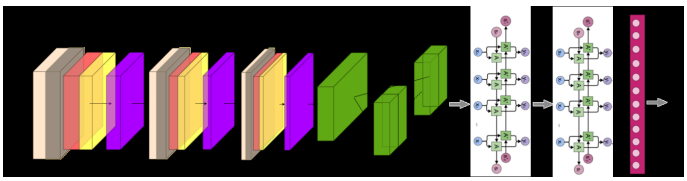

Fig. 4. Schematic representation of CRNN model.[1]

II. Research Work Flow

We first process the music files and derive apt features from them to be able to perform various computations on them for our model. We use the benchmark GTZAN dataset, and extract the Mel Frequency Cepstral Coefficients (MFCC) to represent the audio in time-frequency domain.

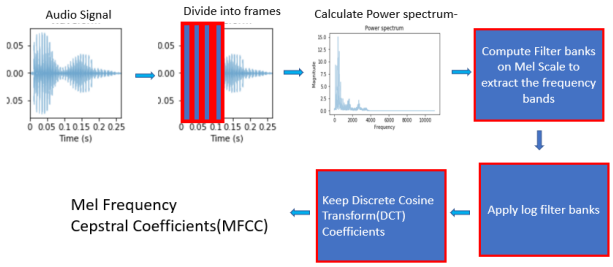

Fig. 5. MFCC derivation.[3]

We divide the dataset into three subsets for training, validation and testing. We train the model using training and validation subsets and then evaluate the performance of our model on the test subset. The CRNN model takes MFCC as inputs and trains using the training and validation sets. After the training, we evaluate out trained model on the test set. We experiment with two different CRNN models: CNN-GRU and CNN-LSTM and achieve better results than CNN [9].

III. Methodology

A. Proposed Hypothesis

Convolutional Recurrent Neural Network model should utilize both spatial and temporal features and produce better results than Convolutional Neural Network model.

B. Dataset Description

For carrying out the task of music genre classification we have used the GTZAN dataset [5] which has been used as a benchmark for vaious researches on music genre classification. The GTZAN dataset provides us 100 music clips each with 30 sec duration for 10 genres each.

C. Mechanism and Model

To get the accurate representation of music signal we use Mel Frequency Cepstral Coefficients (MFCC) which capture the power spectrum of sound. The GTZAN dataset provides us 100 music clips of 30 sec duration for 10 genres each. We process the music clips and derive MFCCs with 23 coefficients. We implement two CRNN models. The models

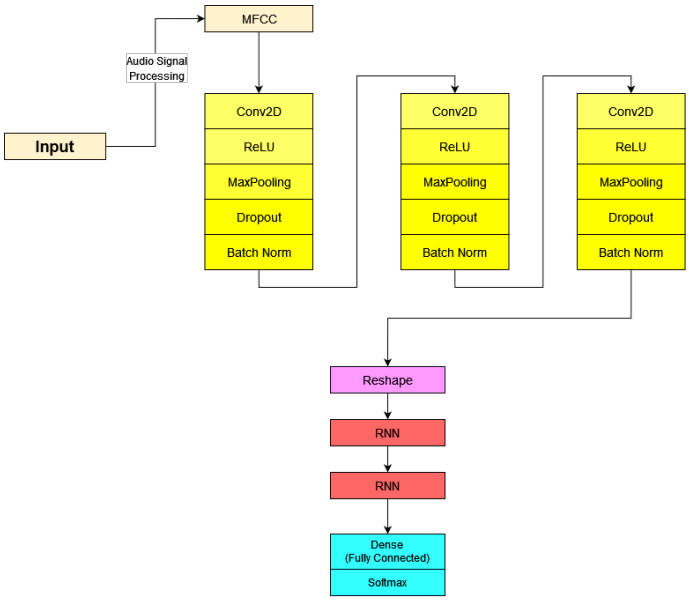

Fig. 6. Architecture of CRNN model

take MFCCs as input. Our models are inspired by [10]. Both our models use 3 Convolutional layers with ReLU activation function where each convolutional layer is followed by a MaxPooling layer, Dropout layer and a BatchNormalization layer. The dropout of these layers is kept low at around 0.1 to avoid overfitting in the RNN layers. For first CRNN model we

have used 2 GRU layers while in our second CRNN model we have used 2 LSTM-RNN layers. However, the output of the Convolutional layers is 3 dimensional whereas the RNN layers need a 2 dimensional input. Therefore, we use a reshape layer to change the dimensions accordingly. We follow the RNN layers in both CRNN models with a Dense layer that uses softmax activation function. We use *librosa* for processing the audio data and deriving MFCC features and we used *keras* to implement our models and train them.

IV. Experiments and Results

We use the GTZAN dataset. We first divide the dataset into training and testing set where the size of the testing set is 20%. We then split the training set into training and validation set. The models are built using *keras*. We use

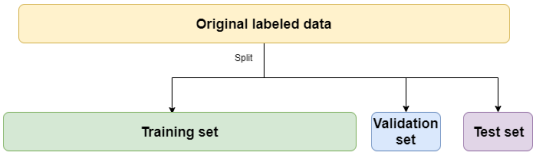

Fig. 7. Schematic representation of splitting dataset.[6]

the *Adam* optimizer for controlling the learning rate while training and *sparse categorical cross entropy* loss function for training. *Adam* optimizer is an extension to stochastic gradient descent and it is based on the adaptive estimation of first order and second order moments. It is a combination of *RMSP* and *gradient descent with momentum* algorithms.

A. Experiment 1

We train and test the CNN-GRU model on our training and test sets.

TABLE I
Parameters for CNN-GRU model

Parameter Name	Parameter Value
Number of Convolution Layers	3
Kernel sizes	(3, 3), (3, 3), (3, 3)
Max Pooling kernel size	(2, 2), (3, 3), (4, 4)
Activation function	ReLU
Dropout after each convolution layer	0.1
Number of GRU layers	2
Number of cells in GRU layers	20, 20

Results:

The CNN-GRU model achieved an accuracy of 85.7%.

B. Experiment 2

We train and test the CNN-LSTM model on our training and test sets.

Results:

The CNN-LSTM model achieved an accuracy of 87.5%.

TABLE II
Parameters for CNN-LSTM model

Parameter Name	Parameter Value
Number of Convolution Layers	3
Kernel sizes	(3, 3), (3, 3), (3, 3)
Max Pooling kernel size	(2, 2), (3, 3), (4, 4)
Activation function	ReLU
Dropout after each convolution layer	0.1
Number of LSTM layers	2
Number of cells in LSTM layers	30, 30

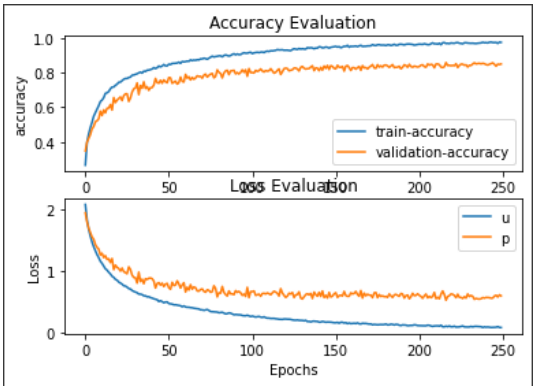

Fig. 8. Training vs Validation plots for accuracies and losses for CNN-GRU

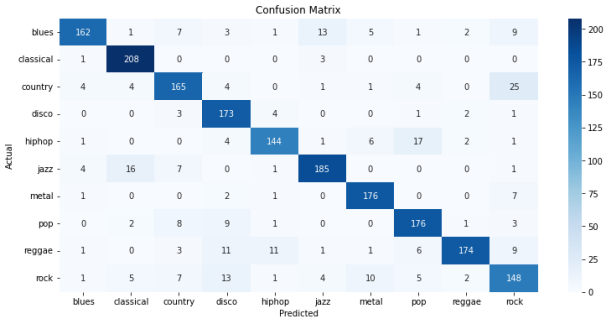

Fig. 9. Confusion matrix for CNN-GRU model

TABLE III
Precision, Recall and F1 scores of CNN-GRU model on the genres.

Genre	Precision	Recall	F1-Score
Blues	0.93	0.79	0.85
Classical	0.88	0.98	0.93
Country	0.82	0.79	0.81
Disco	0.79	0.94	0.86
Hiphop	0.88	0.82	0.85
Jazz	0.89	0.86	0.88
Metal	0.88	0.94	0.91
Pop	0.84	0.88	0.86
Reggae	0.95	0.80	0.87
Rock	0.73	0.76	0.74

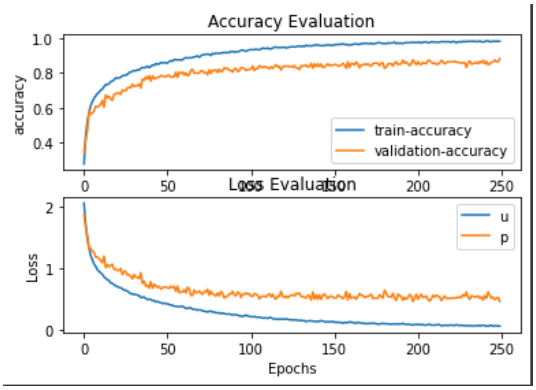

Fig. 10. Training vs Validation plots for accuracies and losses for CNN-LSTM

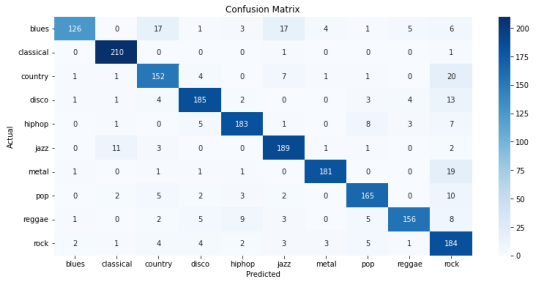

Fig. 11. Confusion matrix for CNN-LSTM model

TABLE IV
PRECISION, RECALL AND F1 SCORES OF CNN-LSTM MODEL ON THE GENRES.

Genre	Precision	Recall	F1-Score
Blues	0.95	0.70	0.81
Classical	0.93	0.99	0.96
Country	0.81	0.81	0.81
Disco	0.89	0.87	0.88
Hiphop	0.90	0.88	0.89
Jazz	0.85	0.91	0.88
Metal	0.95	0.89	0.92
Pop	0.87	0.87	0.87
Reggae	0.92	0.83	0.87
Rock	0.68	0.88	0.77

V. CONCLUSION

We tested two CRNN models for the task of music genre classification to utilize both the hierarchical spatial features and the temporal relationships of the music data. Both the CRNN models outperformed the CNN [9] model and yielded better testing accuracy on the GTZAN dataset. Among the two CRNN models the CNN-LSTM model achieved better results than the CNN-GRU model.

VI. FUTURE SCOPE

We could try to improve the model by using Graph Convolutional Networks to try to capture the relationships and variations of different genres. We could also try to use better datasets and better features to improve the model.

TABLE V
COMPARISON TABLE TO COMPARE THE RESULTS OF OUR MODELS AND OTHER MODELS

Model	Accuracy (Percentage)
Cheng et. al [9]	83.3
de Sousa et. al [16]	79.7
CNN-GRU	85.7
CNN-LSTM	87.5

REFERENCES

[1] An approach towards convolutional recurrent neural networks, available: https://towardsdatascience.com/an-approach-towards-convolutional-recurrent-neural- networks-a2e6ce722b19
[2] Basic concepts of cnn, available: https://medium.com/s-a-a-s/dl-basic-concept-of-cnn- 2ef4fc9b039b.
[3] Deep learning for audio classification, available: https://medium.com/analytics- vidhya/deep-learning-audio-classification-fcbed546a2dd.
[4] Explaining pooling layers, available: https://androidkt.com/explain-pooling-layers- max-pooling-average-pooling-global-average-pooling-and-global-max-pooling/.
[5] Gtzan dataset, available: https://www.kaggle.com/datasets/andradaolteanu-/gtzan- dataset-music-genre-classification.
[6] Test, train and validation split, available: https://www.brainstobytes.com/test- training-and-validation-sets/.
[7] Understanding lstms, available: https://colah.github.io/posts/2015-08-understanding- lstms/.
[8] Ahmet Melih Ba sbu g and Mustafa Sert. Analysis of deep neural network models for acous- tic scene classification. In 2019 27th Signal Processing and Communications Applications Conference (SIU), pages 1–4, 2019.
[9] Yu-Huei Cheng, Pang-Ching Chang, and Che-Nan Kuo. Convolutional neural networks approach for music genre classification. In 2020 International Symposium on Computer, Consumer and Control (IS3C), pages 399–403, 2020.
[10] Keunwoo Choi, Gy orgy Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2392–2396, 2017.
[11] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4):357–366, 1980.
[12] STEVEN B. DAVIS and PAUL MERMELSTEIN. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In Alex Waibel and Kai-Fu Lee, editors, Readings in Speech Recognition, pages 65–74. Morgan Kaufmann, San Francisco, 1990.
[13] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6964–6968, 2014.
[14] Ahmet Elbir, Hilmi Bilal C am, Mehmet Emre Iyican, Berkay Ozt urk, and Nizamettin Aydin. Music genre classification and recommendation by using machine learning techniques. In 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), pages 1–5, 2018.
[15] Prasenjeet Fulzele, Rajat Singh, Naman Kaushik, and Kavita Pandey. A hybrid model for music genre classification using lstm and svm. In 2018 Eleventh International Conference on Contemporary Computing (IC3), pages 1–3, 2018.
[16] Jefferson Martins de Sousa, Eanes Torres Pereira, and Luciana Ribeiro Veloso. A robust music genre classification approach for global and regional music datasets evaluation. In 2016 IEEE International Conference on Digital Signal Processing (DSP), pages 109–113, 2016.
[17] Siddharth Sigtia and Simon Dixon. Improved music feature learning with deep neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6959–6963, 2014.

[18] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302, 2002.

[19] Eve Zheng, Melody Moh, and Teng-Sheng Moh. Music genre classification: A n-gram based musicological approach. In 2017 IEEE 7th International Advance Computing Conference (IACC), pages 671–677, 2017.